PadhAI: Vanishing and Exploding Gradients

One Fourth Labs

Revisiting Gradients:

This particular video focuses on revisiting how the generic formula for the Loss Function in context of RNNs is derived.

$$\frac{\partial \mathscr{L}_t(\theta)}{\partial W} = \frac{\partial \mathscr{L}_t(\theta)}{\partial s_t} \sum_{k=1}^{t} \frac{\partial s_t}{\partial s_k} \frac{\partial s_k}{\partial W}$$

The above formula was derived as we had multiple paths, leading to **W** from **L**, which means we'll have to summate all those previous states to compute the gradients with respect to **W**.

As we know, each of the subsequent states **St** depends on all the previous states **St-1 , St-2, ... , S1**.
Taking in consideration all these states, we can simply write the term in red as:

$$\frac{\partial s_t}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial s_{t-2}} \frac{\partial s_{t-2}}{\partial s_{t-3}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$

We can simplify the equation as :

$$\frac{\partial s_t}{\partial s_k} = \prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}}$$

NOTE : The ***k+1*** here in the product statement shows that we have a total of **t-*k+1*** elements.

Zooming into one element of the chain rule (Part-1):

Some key points till now:

a. If the difference between **t** and ***k*** is very large, the product would have many recurrent gradient terms.

$$\frac{\partial s_t}{\partial s_k} = \prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}}$$

b. We can observe that every pre-activation **(a_j)** is a function of the sigmoid of the previous state. Whereas, the activated state **(s_j)** depends on the pre-activation state of current timestamp only.

$$a_j = Ux_j + Ws_{j-1} + b$$
$$s_j = \sigma(a_j)$$

Based on the statement above, we can simplify one of the factors as :

$$\frac{\partial s_j}{\partial s_{j-1}} = \frac{\partial s_j}{\partial a_j}\frac{\partial a_j}{\partial s_{j-1}}$$

As both the quantities (s_j) and (a_j) are vectors of size R^d , It can be stated that the quantity ds_j / da_j will be a resultant matrix of size R^{dxd}

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \cdots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \ddots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$

So now, talking about one of our previous conclusions, that each post activation state (s_j) depends only on its corresponding pre-activation state (a_j). We can assert that the rest of the terms except the diagonal in the matrix ds_j / da_j will become zero.
Thus, the updated form of the matrix will be:

$$\begin{bmatrix} \sigma'(a_{j1}) & 0 & 0 & 0 \\ 0 & \sigma'(a_{j2}) & 0 & 0 \\ 0 & 0 & \ddots & \\ 0 & 0 & \cdots & \sigma'(a_{jd}) \end{bmatrix}$$

Which can also be written as a simple vector of size R^d:

$$diag(\sigma'(a_j))$$

Zooming into one element of the chain rule (Part-2):

Now, if we take a look at the other element of the chain rule, i.e.,
da_j / ds_{j-1} we can simply say that it would be a sum of the derivative of three terms in a_j w.r.t s_{j-1} as shown below:

$$a_j = Ux_j + Ws_{j-1} + b$$

As the terms Ux_j and b are independent of s_{j-1} the gradient of these terms would also become zero. So now, we can say that :

$$a_j = \cancel{Ux_j} + Ws_{j-1} + \cancel{b}$$

The gradient of the 2nd term which is left to be non-zero, can simply be stated as **W**.
Taking in consideration, a second perspective we can see that da_j / ds_{j-1} will be of size R^{dxd} and so is the matrix **W**.

So now, we can write the whole elements which was earlier divide into two products to be:

$$\frac{\partial s_j}{\partial s_{j-1}} = diag(\sigma'(a_j))W$$

According to our update rule in Gradient Descent Algorithm, the problem of Vanishing gradients occurs when the update term is too small, that the change would be too small.
On the other hand, exploding gradients is the problem that occurs whenever the update term is so large that the update lets it overshoots the minima.

A small detour to calculus:

Now, if we want to know what would be the magnitude of the given term, let's have a look at the basic calculus.
If we want to minimize or maximize a function w.r.t **x**, i.e we want to find the value of **x**, such that the first derivative of function converges down to **zero**.
If we want to know whether the particular value of **x** gives us the minimum or maximum value for that particular function, we need to compute the second order derivative of the same function.
If the second order derivative of that function is positive, the the value of **x**, for which the first derivative is **zero**.

Example:
Let $f(x) = x^2 + 2x$
Putting $df(x)/dx = 0$, we get $x = -1$.
To know whether the function will be taking either the maximum or the minimum value at $x = -1$.

We'll have to check whether the second order derivative $d^2f(x) / dx^2$ will be a positive or negative quantity.
As, $d^2f(x) / dx^2 = 2$ is a positive value, we can say that the value of function $f(x)$ at $x = -1$ is minimum.

Looking at the magnitude of the derivative:

According to the Cauchy-Schwarz Inequality theorem, we can state the following:

||a*b|| <= ||a||*||b||

Note: Refer to the following link for a proof
https://www.khanacademy.org/math/linear-algebra/vectors-and-spaces/dot-cross-products/v/proof-of-the-cauchy-schwarz-inequality.

When we particularly talk about the ***sigmoid*** non-linearity, as we've seen in the lecture that it takes the maximum value for the function **x*(1-x)** at **x = ½**.

When we talk about RNNs, ***tanh*** non-linearity is often been used, and the maximum value for the function $\mathbf{1-x^2}$ is at **x = 0**.

Now that, the maximum values for both the functions are:

$$\begin{aligned}\sigma'(a_j) &\leq \frac{1}{4} = \gamma \ [if\ \sigma\ is\ logistic] \\ &\leq 1 = \gamma \ [if\ \sigma\ is\ tanh]\end{aligned}$$

Arriving at a specific term, we have:

$$||\frac{\partial s_j}{\partial s_{j-1}}|| \leq \gamma ||W|| \leq \gamma\lambda$$

Exploding and Vanishing Gradients:

As per the concept, we know what Vanishing and Exploding gradient actually means.

If we take a look at the earlier derived **product**, which consists of several similar derivatives:

$$\begin{aligned}\left\|\frac{\partial s_t}{\partial s_k}\right\| &= \left\|\prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}}\right\| \\ &\leq \prod_{j=k+1}^{t} \gamma\lambda \\ &\leq (\gamma\lambda)^{t-k}\end{aligned}$$

So, if the value for ***(gamma*lambda)*** is smaller than 1, and ***(t-k)*** is very large, the resultant would be a really smaller quantity. Which will be termed as a **vanishing gradients** problem
Similarly, if the resultant value is larger enough, the problem here would be **exploding gradients**.